Literacy rules

Verhasselt, Els

Chapter Four

Methodology

At macro level, conducting cross-national comparative studies, are the paramount goal of the social sciences. They enable us to know whether observed social phenomena and relationships are universal or confined to a particular nation or type of society. They lend themselves to compare social groups. Such studies are rife with potential problems of item and sample comparability. These matters, along with survey methodology, data collection and processing, instrument validity, and the data quality of the IALS as a whole, will therefore be discussed firstly in section 4.1.

Estimates from sample surveys are always subject to sampling errors since data are obtained from only a portion of the population. Generally, sampling errors can be easily controlled and measured in probability samples. However, it does not include any additional errors that may occur because of the practical difficulties involved in conducting a survey, such as subtle differences in design and implementation and in the pattern of non-response across countries, which, if present, can lead to overestimation or underestimation of the true size of differences between populations. Statistics Canada, ETS and the national study teams have performed extensive analyses to understand the nature and extent of error associated with the differences in design and implementation. Measures used in the IALS to counter both types of error are presented.

In addition, the IALS has been the subject of an independent quality review. Furthermore, a large number of guidelines, technical specifications and other documents have been written and made available to the national study teams in the participating countries. In sum, data quality has been treated as key issue.

A next section introduces the dissertation methodology to compare Flanders with the Netherlands within the IALS setting. Log-linear models will be used for this purpose. Therefore, some basic notions on log-linear models are described: the broader contingency table analysis perspective, specific terminology as there are odds and odds ratio's, saturated and unsaturated models, frequency and logit models, dummy and effect coding and the very interpretation methods.

4.1. The IALS : Survey Administration, Response And Data Quality

Information on these topics, described in this section, was found in the international IALS reports '*Literacy, Economy and Society'* and '*Literacy Skills for the knowledge society'* and more elaborately in '*Adult Literacy in OECD Countries – Technical Report on the First International Adult Literacy Survey'* (Murray et al., 1998). In addition, the IALS Micro-data Package and the Flemish report (Van Damme et al., 1997) were consulted. Although improvements can be made in some aspects of the survey methodology used, the International Adult Literacy Survey made an important first step in the area of International Survey Research.

4.1.1. SURVEY METHODOLOGY

Comparability in the measurement instruments is a necessary but insufficient condition for the comparability of data in multi-national surveys. To fully evaluate data comparability, consideration must also be given to other survey design aspects.

4.1.1.1. Target Population And Frame Coverage

Each country participating in the IALS designed a probability sample from which results representative of the civilian non-institutionalised population aged 16 to 65 could be derived. To draw a probability sample needs knowledge of the chance of an individual entering the sample. This chance cannot be zero for no one. In this way, population parameters can be measured without any bias.

All samples excluded full-time members of the military and people residing in institutions such as prisons, hospitals, and psychiatric facilities. Moreover, the survey was carried out in the national language. When respondents could not speak the designated language, attempts were made to complete the background questionnaire, so that their literacy level could be estimated and the risk of distorted results would be reduced.

Particularly, Belgium also excluded the Flemish inhabitants of the Brussels region. Because the Belgium IALS-sample is representative of the "Flemish Region", excluding Brussels, the name Flanders is used throughout, instead of the more conventional term "Flemish Community". In addition, countries were permitted to sample abroad age-categories: the Netherlands were especially interested in the older adults aged 65 to 74, while Flanders gave interest in the school leavers aged 17. Finally, the total number of exclusions was estimated at less than 1 percent, for both Flanders and the Netherlands. The Flemish total-in-scope population was 3,692,116 persons and for the Netherlands 11,495,719.

4.1.1.2. Sample Design

Initially, each IALS country was required to provide a sample of 3,000 survey respondents per test language. Yet, several countries were unable to secure sufficient financial resources to support such an expensive undertaking (Darcovich, 1998, p.25). Therefore, the final sample design criteria specified at least 1,000 respondents, which ensured sufficient information for the calculation of reliable literacy profiles. No uniform sampling plan was imposed due to differences in the data sources, collection practices and resources available in each of the participating countries.

The designated area of *Flanders* was divided by the National Institute for Statistics (NIS - Nationaal Instituut voor Statistiek) into 7203 statistical sectors, from which 200 were selected with probability proportional to size. Then, 40 persons were chosen from a

complete list of people for each of the selected sectors. All selected people were sent a letter informing that an interviewer would be visiting to conduct a literacy survey.
Finally, as low educated people were expected to be underrepresented in IALS-like research, special attention had to be taken in order to get an equal distribution of people by education level. Unlike most other participating countries, Flanders did not have any information on education available at population level, due to the constitutional privacy law. Therefore, a more complex method was used in drawing a sample, following the specifications given by Nancy Darcovich of ETS. Reflecting the Flemish (from 16- till 65-year-olds) distribution of educational attainment, the sample size was divided into three education levels, that is 'high', 'middle', and 'low, respectively containing 17%, 30%, and 53% of the total Flemish adult population. Then, an administrative selection procedure was inserted: (1) 'low' - all low-educated people who wanted to engage in the IALS were consequently interviewed; (2) 'middle' - half of the respondents who completed higher secondary education and (3) 'high' - two thirds of the higher educated, who were willing to cooperate in the IALS, were withdrawn from the sample and were given a short questionnaire, from which the results are not included in the sample. The resulting number of respondents in the final sample was 2,261 people.

The Dutch approach was to use a two-stage systematic sampling procedure. In the first stage, 9,000 (7,000 and 2,000 in reserve) postal codes were selected from a NIPO data file (Dutch National Centre for Public Opinion and Marketing Research - Nederlands Instituut voor de Publieke Opinie en het Marktonderzoek) of 540,817 codes, including information on educational attainment. In the second stage, one address was chosen from each selected postal code. All selected addresses were mailed a letter informing the household that an interviewer would be visiting to conduct a literacy survey. Within the selected households, the person with the first-occurring birthday in the year was selected to participate in the survey. A total of 3,090 responses was obtained.

4.1.1.3. Weighting And Benchmarking

Sample weights are used to compensate for non-response, non-coverage, and for the use of unequal selection probabilities. Developing weights involves correction for differential response rates within classes of the sample and adjustment of the sample distribution by demographic variables to known population distributions, from non-IALS sources. This 'benchmarking' procedure, as a final step in the weighting process, assumes that the characteristics of the respondents are similar to those of the non-respondents. The primary purpose of weighting adjustments is to reduce bias in the survey estimates. However, large variations in weights can seriously inflate the variances of the estimates.

In Flanders, counts were used from the 1991 Census as the benchmarking variables: region, age, gender and education. In the Netherlands, each respondent was assigned a 'base weight'. This weight was calculated by dividing the in-scope target population size by the number of respondents. The Netherlands, then, adjusted its sample counts to make them correspond to the 1994 counts organised by the Central Bureau of Statistics (CBS) on Dutch inhabitants aged 16 to 74. These adjustments were based on four demographic characteristics: region, age, gender and education.

4.1.2. Data Collection And Processing

The quality of the IALS data depends on the data collection procedures used. Important factors include the experience and the training of the interviewers, their supervision, the quality checks performed on their work, the way in which the survey is introduced to the

sampled persons, and the procedures used during the interviews. In addition, the coding needs to be performed in a consistent way across countries. Errors can occur at this stage, as well as while entering the data in computer files.

As described earlier, the IALS gathered descriptive and proficiency information through a background questionnaire and a literacy assessment. Survey respondents spent approximately 30 minutes answering a common set of background questions. Responses to these questions make it possible to summarise the survey results by means of an array of descriptive variables, and also increase the accuracy of the proficiency estimates for various sub-populations. Background information was collected by trained interviewers at the respondents' homes.
Once the background questionnaire was completed, the interviewer presented a core booklet containing six simple literacy tasks, designed to avoid the embarrassment of giving the full test to adults with very low literacy skills. Only those able to complete at least two of these correctly were given the full test with a larger variety of tasks. Most of these tasks were open-ended requiring of the respondents to provide a written answer.

4.1.2.1. Model Procedures

Each IALS country was given a set of administration manuals and survey instruments to use as a model. Countries were authorised to adapt these models to their own national data collection systems, but they were required to retain a number of key features. Firstly, respondents were to complete the core and main test booklets alone, at their home, without any help from another person or without using a calculator. Secondly, respondents were not to be given monetary incentives to participate. Thirdly, despite the prohibition against monetary incentives, interviewers could follow procedures to maximise the number of completed background questionnaires, and were to use a common set of coding specifications to deal with non-response. The latter requirement is critical. Because non-completion of the core and main task booklets is correlated with ability, background information about non-respondents is needed to impute cognitive data to these people.

4.1.2.2. Reducing Non-Response Bias

Besides the rules governing the background questionnaires and the item modification process, a crucial step contains the managing of non-respondent cases in a uniform manner, in order to limit the level of non-response bias in the resulting survey estimates. In the IALS, the respondent burden plays a substantial role because the IALS contents may be threatening to some respondents. Non-response rates are so high that bias in the survey estimates is inevitable. More precisely, a respondent had to complete the background questionnaire, pass the core block of literacy tasks, and attempt at least five tasks per literacy scale in order for researchers to be able to estimate his or her literacy skills directly. Literacy proficiency data were *imputed* to individuals who failed or refused to perform the core literacy tasks and to those who passed the core block but did not attempt at least five tasks per literacy scale. In sum, the definition of a respondent in the IALS is a person who partially or fully completed the background questionnaire. In some cases, incomplete assessment data were obtained, but when the individual provided background information and indicated why he or she did not complete the core and main literacy task booklets, it was possible to impute a literacy profile to that person. Formulating secure definitions of response versus non-response is essential to limit non-response bias. In addition, *precautions* were made to reduce non-response bias. Interviewers were instructed to call back households that were difficult to contact. They were also given a detailed non-response

classification to use in the survey. And, all countries' sample designs included some over-sampling.

Finally, an analysis was conducted to study potential non-response bias in IALS. Firstly, the characteristics of the non-respondents were examined whether one group in particular was not responding to the survey. This issue is a matter of concern when such a group is defined by a characteristic that is strongly correlated with literacy level. Therefore, the IALS sampling guidelines included an adjustment during the weighting procedure. This adjustment, known as *post-stratification*, offers protection against extreme sample configurations in adjusting the population weights so that they match known population parameters, e.g. by age group or education level. Secondly, the weights before and after the post-stratification adjustment were also compared with each other. An evaluation of the post-stratification model was included in the analyses. Thirdly, the level of bias that would have had to be present to alter the IALS estimates significantly, was calculated. The results show that the levels of the bias that may be present are too low to alter any of the main IALS findings. Exhaustive non-response bias analyses can be found in Murray *et al*. (1998).

4.1.2.3. Survey Response

None of the countries met the non-response rate requirements which were less than 20 percent overall and for all important strata. As stated in the IALS reports, the response rate is calculated as the ratio of the response rate and the total survey sample without the out-of-scope respondents. Flanders has the lowest response rate of all participating IALS countries: 2,261 respondents covering 36 percent (2,261/[(8,880-2,667)]*100). The Netherlands have 45 percent response rate (3,090/[(9,000-2,099)]*100). In general, though, it should be noted that both the Flemish and Dutch surveys often have high non-response rates, mainly because a large number of interviews are conducted in the country on an annual basis. Experience shows that social surveys frequently encounter serious response problems.
However, the IALS survey response being looked into more elaborately, the out-of-scope population is not equally determined. The sampling procedure in the Netherlands by which the addresses are selected from a postal code region defines the out-of-scope group partially as addresses of business and institutions and partially as not belonging to the target age group. In Flanders, the 2,667 out-of-scope respondents are those who were willing to commit themselves to the IALS but had to fill in a shorter questionnaire because their level of educational attainment was defined as 'middle' or as 'high'. This administrative selection procedure, fostered by the specific sampling situation in Flanders and unique for the IALS as a whole, is mainly responsible for the low response rate in Flanders.

Table 4.1: The IALS response rate in Flanders and the Netherlands

Survey	Flanders	The Netherlands
Response	2,261	3,090
Non-response	3,952	3,811
Out of Scope	2,667	2,099
TOTAL	8,880	9,000

4.1.2.4. Scoring

Unlike multiple-choice questions, open-ended items such as those used in the IALS elicit a large variety of responses. Because raw data are seldom useful by themselves, responses must be grouped in some way in order to summarise the performance results. According to

their scores, responses to the IALS open-ended items were classified as correct, incorrect, or omitted. The models employed to estimate ability and difficulty are predicated on the assumption that the scoring rubrics developed for the assessment were applied in a proper and consistent way within and between countries. Once the tests are scored and the literacy profiles are calibrated, an analysis of the model parameters and the psychometric functioning of the tests can reveal whether the tests behaved similarly within and across countries.

4.1.2.4.1. Intra-Country Rescore Reliability

A variable sampling ratio procedure was set up to monitor scoring accuracy. At the beginning of the scoring, almost all responses were rescored to identify inaccurate scorers and to detect unique or difficult responses that were not covered in the scoring manual. The rescoring ratio was brought to a maintenance level to monitor the accuracy of all scorers, after a satisfactory level of accuracy was achieved. Within each country, at least 20 percent of the tests were required to be rescored. Average agreements were calculated across all items. To ensure that the first and second scores were truly independent, certain precautions had to be taken. Scorers, for example, had to be different persons, and the second scorer was not allowed to see the scores given by the first scorer. The performance of the scorers who received identical training within a country was expected to be more consistent with one another than with the performance of the scorers in other countries. This expectation was confirmed: most of the rescoring reliabilities were above 97 percent. It is important to note that these results lay well within the statistical tolerance set by the IALS study, i.e. a match of the two sets of scores with at least 95 percent accuracy, and are considerably better than those realised in other large-scale studies using open-ended items.

After intra-country reliabilities were calculated, a few scorers were found to be unreliable. These scorers either received additional training or were excluded. Whenever scores and rescores differed, the first scores were replaced with correct scores if inaccuracy was due to a systematic error on the part of the first scorer. Furthermore, in some cases the scoring guide was found to be ambiguous. Then, the scoring guide was revised and the first scores were amended to reflect the revisions, while the second scores were not altered. In sum, the first scores reflect changes and corrections resulting from lessons learned from the intra-country rescoring analysis. However, the extent to which the reliabilities are underestimated must be very small, given that most of the reliability lie above 97 percent. These values indicate that very consistent scoring was achieved by all participating countries.

4.1.2.4.2. Inter-Country Rescore Reliability

Even after ensuring that all scorers were scoring consistently, fixing ambiguities in the scoring guides, and correcting any systematic scoring errors, it was still necessary to examine the comparability of scores across countries. Accurate and consistent scoring within a country does not necessarily imply that all countries are applying the scoring guidelines in the same manner. Scoring bias may be introduced if one country has responses that are scored differently from the other countries. The inter-country rescorings were undertaken to ensure scoring comparability across countries: each country had 10 percent of their sample rescored by the scorers in another participating country. Inter-country score reliability was calculated by Statistics Canada and later evaluated by ETS. Every country was required to introduce a few minor changes to the scoring procedures whereby relying on the evaluation. Using the inter-country score reliabilities, researchers can identify poorly constructed items, ambiguous scoring criteria, erroneous translations of items or scoring criteria, erroneous printing of items or scoring criteria, scorer inaccuracy, and, most importantly, situations in

which one country consistently scored differently from another. In the latter circumstance, scorers from one country may consistently rate a certain response as being correct while those from another country regard the same response as incorrect. This type of score asymmetry must be eliminated before the IRT scaling is performed. ETS and Statistics Canada identified such items, while the country in which the scoring problem occurred investigated the plausible causes of such systematic score bias. Whenever a systematic error was identified in a particular country, the original scores for that item were corrected throughout the entire sample.

4.1.2.4.3. Inter- And Intra-Country Rescore Reliability In Flanders And The Netherlands

In practice, this rescoring stage evolves in a rather rickety way caused by several circumstances. The Netherlands carried out both inter- and intra-country rescoring internally due to a lack of available language experts in Dutch. Separate groups were established to perform the rescoring. As for Flanders, as it can be read in the international and national IALS report, 300 booklets were rescored by the Netherlands, resulting in an average agreement of 94 percent, while no asymmetric items were found. However, this average agreement can be considered as being a rather moderate result. Indeed, the Dutch researchers who coded the IALS items in the first round of the IALS in 1994, were not available anymore. Therefore, the rescoring was performed by two Dutch junior researchers who also were trained by the Flemish national study manager. In sum, one may give reasonable doubt to the seriousness of the inter-country rescore reliability for Flanders and the Netherlands.

4.1.2.5. Data Capture, Data Processing and Coding

As a condition for their participation in the IALS, countries were required to capture and process their files by using procedures that ensured logical consistency and acceptable levels of data capture error. More specifically, countries were advised to conduct complete verification of the captured scores by entering each record twice in order to minimise error rates. Because the process of capturing the test scores in an accurate way is essential to high data quality, 100 percent keystroke validation was needed.

Each country was also responsible for coding industry, occupation, and education using standard international coding schemes (International Standard Industrial Classification, or ISIC; International Standard Occupational Classification, or ISOC; and International Standard Classification of Education, or ISCED). Further, coding schemes were provided for open-ended items, and countries were given specific instructions about the coding of such items so that coding error could be contained to acceptable levels.

Furthermore, to create a workable comparative analysis, each IALS country was required to map its national dataset into a highly structured, standardised record layout. In addition to specifying the position, format, and length of each field, this International Record Layout included a description of each variable and indicated the categories and codes to be provided for that variable. Each country was instructed to perform a 100% verification of the background questionnaire data entry and of the entry of strings of test scores. Upon receiving a country's file, Statistics Canada performed a series of range checks to ensure compliance to the prescribed format. When anomalies were detected, countries corrected the problems and submitted new files.

4.1.3. DATA QUALITY

4.1.3.1. IALS Quality Level Guidelines

Because within all survey data the data quality is affected by both sampling and non-sampling errors, IALS describes some quality level guidelines. Basically, in measuring the potential size of sampling errors its fundament is the standard error of the estimates derived from survey results. Therefore, data quality levels in IALS are determined by (1) the number of respondents who contribute to the calculation of the estimate and (2) the coefficient of variation (CV) being the standard error of an estimate expressed as a percentage of that estimate. Data quality levels are acceptable, marginal or unacceptable.
Firstly, if the number of respondents who contribute to the calculation of the estimate is less than 30, the weighted estimate should be considered to be of unacceptable quality. Secondly, for weighted estimates based on sample sizes of 30 or more, users should determine the coefficient of variation of the estimate. If this coefficient varies between 16.6% and 33.3%, the quality level of the estimate is marginal; if this coefficient is low and varies between 0% and 16.5%, the estimate is of acceptable quality.

In addition, because the IALS surveys are based upon complex sample designs, with stratification, multiple stages of selection, and unequal probabilities of selection of respondents, other aspects have to be taken in account. These data present problems since the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used. In order for survey estimates and analyses to be free from bias, the survey weights must be used.

Concluding, taking these quality level guidelines seriously, the IALS considers all estimates to be releasable; those of marginal or unacceptable quality level will be -following the IALS Release Guidelines- accompanied by a warning.

4.1.3.2. Independent Quality Review

When new kinds of assessments are undertaken, like the IALS, some concerns, comments, and critics raise, both by participating countries during the several IALS stages and by others after the results have been published internationally. In these, the concerns of France have to be mentioned; France joined the IALS at the earliest stage and its experts collaborated with representatives from the other countries to design the assessment instruments and conduct the field test and main data collection. In august 1995, when the international results became available for scrutiny, the French authorities began to question the comparability of the results between countries, they questioned the appropriateness of the assessment instruments, the validity of the sampling procedures and the reliability of the population estimates.
To address these concerns, Statistics Canada decided to subject the IALS methodology to an external evaluation. Three survey methodologists who have had no previous involvement in any phase of the IALS undertook an independent review. Due to a lack of time and resources, they were not able to analyse the survey data. The approach that has been taken was collecting information on the survey procedures employed by the eight countries participating in the first round of the IALS. A questionnaire covering the important methodological issues was developed and completed –in most cases- by one of the reviewers on a visit to the national survey organisation.

Due to these arrangements, the report can only point out where survey procedural errors occurred and does not investigate what the likely consequences of such errors might be. Therefore, the following recommendations are made:

1. Problems arise in most design components across countries, especially with regard to sample design and non-response. These problems threaten the validity of any comparison of literacy levels across countries. It is strongly recommended not to publish tables that rank countries by literacy level.
2. Nevertheless, the IALS collected valuable data that should be published, focussing on the correlates of literacy in the different countries and on the comparison of these correlates across these countries. Notwithstanding, this matter should be treated with caution.
3. Further investigation is needed to assess the quality of the data, i.e. non-response, weighting and benchmarking schemes.
4. Future IALS rounds need standardisation of survey procedures across countries. Careful monitoring is required, which will have cost implications.

Yet, despite these actions, a fierce controversy arose with the French authorities deciding in October 1995 to withdraw their country's results, after it emerged that it had the lowest scores on all three domains. In addition, the European Commission set up a project to re-evaluate the IALS results, the results of which are presented in *Measuring Adult Literacy – The International Adult Literacy Survey in the European context* (ONS, 2000).

4.1.4. INSTRUMENT VALIDITY

Given the accuracy and consistency of the scoring, the test had to perform similarly in each country, as another condition for the success of the project. The IALS test items needed to pose the same level of difficulty regardless of the respondent's background or country of origin. Instrument validity has already been discussed in Chapter Three.

4.2. Comparing Flanders And The Netherlands By Using Log-linear Models

The advantage of statistical models that summarise data and test hypotheses is well recognised. The standard regression model (and path model as its extension), which assumes linear relations between variables measured at interval level without any interaction effect, as well as a number of restrictions about the error term that has to justify the use of ordinary least squares estimation, is just not applicable to categorical data measured at nominal or ordinal level. Variants of this standard model sometimes offer solutions. However, despite these difficulties, statistical tests in regression analyses are usually based on the assumption of underlying normal distributions, at least of the error terms. Most of the times, this assumption is very unrealistic with regard to categorical variables. Indeed, table analysis appears to be a statistical method closest to the cross-sectional nature of IALS-data.

In the early 1970s, the log-linear model (Goodman [1970-1973] in: Hagenaars, 1993) was introduced into social science research on contingency table analysis of categorical data. This model is useful for uncovering the potentially complex relationships between the variables in a multi-way cross-tabulation and provides a natural way of dealing with them. Indeed, it provides better and more appropriate test statistics; it provides a global test for each independent variable rather than a series of tests as was the case for cross-tabular analysis. In addition, the results allow for a test of main effects separate from interaction

effects. Moreover, log-linear analysis surpasses cross-tabulation by simultaneously estimating the effects of multiple variables (Kaufman & Schervish, 1986, p.717).

In addition, and as stated in the Sage Publication Series (Lee et al., 1989) on Quantitative Applications in the Social Sciences, the log-linear approach has the following advantages. This method can be applied to large contingency tables. The sample size requirements are not that stringent, because they only focus on the marginal totals. Yet, considering the type of problem that can be addressed by the procedure, a burden remains because the procedure is limited to the log-linear hierarchical framework (see 4.2.1.4.). In addition, the log-linear model approach is a replication-based method of variance calculation adapted to the maximum likelihood estimation.

It should not be surprising that since the 1970s the log-linear model has become the dominant form of categorical data analysis. It has strengthened its position as more and more social investigators have fruitfully applied it to their research (Hagenaars, 1990).

4.2.1. LOG-LINEAR MODELS : AN INTRODUCTION

Hagenaars (1990) presented log-linear modelling in an exhaustive and clear way so it became easy to comprehend. Therefore, and because of pragmatic considerations, this introduction into log-linear modelling and the description of the use and interpretation of log-linear estimates as a whole is generally based on what Hagenaars elaborated on in his book.

To begin with, the starting point of the analysis of the relations between categorical variables is a multivariate frequency table. As the number of cases in each cell can be expressed as the product of a number of parameters, including interaction factors, the underlying model is called a multiplicative model. To obtain an additive (linear) model, the equation of which consists of terms and not of factors – and thus is more easy to comprehend -, a logarithmic transformation is needed, by using the natural logarithm of the equation.
Though, before going into the log-linear model itself, some basic preliminary concepts are introduced here, such as 'maximum likelihood estimation', 'odds' and 'odds ratios', 'saturated' and 'unsaturated' models, 'frequency' and 'logit' models, and 'dummy' and 'effect' coding. In addition, a section deals with the notation in modelled equations.

4.2.1.1. Maximum Likelihood Estimation

The *maximum likelihood principle* is a statistical principle applied to find 'good' parameters. Hagenaars uses the description of Hays (1981) who describes this principle in a concise but intuitively very clear way: "In effect this principle says that when faced with several parameter values, any of which might be the true one for a population, the best "bet" is that parameter value which would have made the sample actually obtained have the highest prior probability. When in doubt, place your bet on that parameter value which would have made the obtained result most likely".

4.2.1.2. Odds And Odds Ratios

Log-linear models require a very different way of thinking about relationships between variables from the one which many social scientists are used to. Percentages are deeply ingrained in social research, but log-linear models are non-linear in terms of percentages, and thus the effects they represent can not be viewed adequately in terms of percentage differences.

Odds and *odds ratios* are the concepts through which the log-linear parameters can be interpreted. Odds are defined as the frequency (or probability) of one category of a variable compared to the frequency (or probability) of another. Let us assume that variable A has two subcategories "a" and "not a". This points out that each respondent either has or has not characteristic a. Once a sample is drawn, the probability that an individual randomly selected from the total sample carries this characteristic a, will be known, as well as the probability that this individual will not carry it. Expressing the inequality of these two probabilities by taking their ratio, results into the *odds* of having characteristic a rather than not having a. If the two probabilities are equal, the odds are 1. Because these odds are computed on the basis of the marginal distributions, they are sometimes called *marginal odds*.

Then, a variable B having brought in with the sub-categories "b" and "not b", the odds of having "a" can be calculated for the subgroup respondents who have b and for the subgroup who does not have b. These odds are called *conditional odds*. The more the two conditional odds deviate from each other, the stronger the association becomes between variable A and B. A measure of this deviation is the ratio of the two conditional odds, namely the *odds ratio*. In other words, the odds ratio is a symmetric measure of association and is independent of the marginal distributions of the variables.

Finally, *partial odds* are defined as average conditional odds (the geometric mean). The partial odds of having a is what the odds of having a rather than not having a are on average among those having b and those not having b.

4.2.1.3. Notation

Observed frequencies are indicated by f; superscripts refer to variables and subscripts to categories of these variables. The observed proportion p of individuals belonging to category i of variable A, category j of variable B and category k of variable C, is multiplied by the total sample size, namely N.

$$f^{ABC}_{ijk} = N\, p^{ABC}_{ijk}$$

Probabilities in the population are denoted by π (pi). F represents the frequencies that would have been found if the sample had been an exact reflection of the population without sampling fluctuations.

$$F^{ABC}_{ijk} = N\, \pi^{ABC}_{ijk}$$

Table analysis can be defined as getting estimates of the population parameters π through the use of proportion p departing from certain hypotheses about the population. The relationships in a three dimensional frequency table can be fully described by the following log-linear model, rendered into its multiplicative form with parameters denoted by τ (tau).

$$F^{ABC}_{ijk} = \eta\, \tau^{A}_{i}\, \tau^{B}_{j}\, \tau^{C}_{k}\, \tau^{AB}_{ij}\, \tau^{AC}_{ik}\, \tau^{BC}_{jk}\, \tau^{ABC}_{ijk}$$

One could speak of an overall effect (η - *eta*) on *F*, of three main or one-variable effects (τ^{A}_{i} τ^{B}_{j}, *and* τ^{C}_{k}), three direct or two-variable effects (τ^{AB}_{ij}, τ^{AC}_{ik}, *and* τ^{BC}_{jk}) and one interaction order or three-variable effect (τ^{ABC}_{ijk}*).*

Hence, a multiplicative model is very difficult to comprehend, even through statistical means. An additive model that represents a linear relationship between variables is much easier to study and understand. Therefore, it is more convenient to work with the logarithmic transformation of this latter equation, using θ (theta) and λ (lambda). This is the logarithmic, additive form that has given the log-linear model its name.

$$G^{ABC}_{ijk} = \theta + \lambda^{A}_{i} + \lambda^{B}_{j} + \lambda^{C}_{k} + \lambda^{AB}_{ij} + \lambda^{AC}_{ik} + \lambda^{BC}_{jk} + \lambda^{ABC}_{ijk}$$

4.2.1.4. Saturated And Unsaturated Models

A *saturated model* imposes no restrictions on the relations between the variables. As such, the saturated model fully represents the observed data. It reproduces the observations exactly in terms and number of effects with all possible effects present. *Unsaturated models* reflect the investigator's expectations concerning the data in the form of a priori restrictions on the effect parameters, mostly by assuming that one or more (interaction) effects between variables are absent. The validity of restrictions imposed by an unsaturated model can be tested empirically by setting up the relevant log-linear model, computing the estimated expected frequencies for this model and comparing them with the observed frequencies. If the hypothesis has to be rejected, a next step is to look for a model that does fit the observed data, which is mainly guided by theoretical notions, but also to make use of several statistics which make clear to what extent and at which points particular models deviate from reality. Besides theoretical meaningfulness and test statistics, the leading principle of exploratory research procedures is also the parsimony principle: given the importance of accuracy, a less complex explanation of the data is to be preferred to a more complex one. Simple models provide us with more information about the state of the world, as they exclude more possible states of the world than complex models do.

Further, a distinction can be made between *hierarchical and non-hierarchical models*. A hierarchical model must have lower order terms for all possible combinations of the variables, if a term exists for the interaction within this set of variables.
Unsaturated hierarchical models may be interpreted as models, the information of which is contained in certain marginal frequencies in combination with the a priori restrictions. Moreover, these models are assumed to be sufficient to describe what is presented by the total table. Therefore, such models can be uniquely represented by the reproduced marginal frequencies.

4.2.1.5. Frequency And Logit Models

Some log-linear models do not make any distinction between independent and dependent variables, between causes and effects. Models which treat all variables as belonging to the same causal level are called *frequency models*. Log-linear models that do make the distinction are labelled *effect models* or *logit models*. The effect parameters of a logit model have a causal connotation contrary to the 'effect' parameters of the frequency model. Effect models can be used in a modified multiple regression approach and in a modified path model approach.

4.2.1.6. Dummy Coding And Effect Coding

Log-linear equations contain too many parameters to be identifiable. The fact that effects are defined by comparison, provides a general solution to the identification problem. This solution involves dummy-coding or effect-coding.

Applying *dummy coding* as a way of reducing the number of parameters, can be done by taking a dummy variable and setting the effect of one category equal to an arbitrary constant (1). In this way one can see to what extent the effects of the other category deviate from this arbitrary value. Most applications of log-linear modelling use *effect coding*, though. Effect coding implies the log-linear parameters over any subscript totalling zero and the parameters of the multiplicative model multiplied over any subscript equalling one. Then, log-linear effects are expressed as deviations from the average effect.

Effect-coding and dummy coding are arbitrary parameterisations of basically the same model with the same estimated expected frequencies. They both lead to the same substantive conclusions about the relationships in the multivariate frequency table. Nevertheless, the values of the parameters themselves and their interpretations depend on which parameterisation is used.

4.2.2. LOG-LINEAR MODELS: INTERPRETATION

Although log-linear models have become widely accepted as a tool for analysing categorical data, still their full power is not always exploited because the interpretation of the parameters remains a cumbersome aspect. This section will pay attention to the estimation of the parameters, to the method of interpretation as postulated by Kaufman and Schervish (1986), and to causality in general.

4.2.2.1. The Method Of Kaufman And Schervish For Parameter Estimation

When going back to the log-linear equation postulating the relationships within a three dimensional frequency table, one can distinguish the overall effect η, three one-variable effects, three two-variable effects and one three-variable or second order effect.

$$F^{ABC}_{ijk} = \eta\, \tau^{A}_{i}\, \tau^{B}_{j}\, \tau^{C}_{k}\, \tau^{AB}_{ij}\, \tau^{AC}_{ik}\, \tau^{BC}_{jk}\, \tau^{ABC}_{ijk}$$

The *overall effect* reflects the mean level of all cell frequencies. Moreover, it is merely a reflection of the sample size. If the sample size doubled, all cell frequencies would be twice as high and, accordingly, the estimation of η twice its value.

The *one-variable effects* reflect the fact that cell frequencies are expected to be higher or lower depending on the distribution of the single variables. These effects turn out to be the partial odds, which means that the average probability of being in category A=i is compared with the overall probability of being in any of the table cells. As for a dichotomous variable, the ratio of the two partial odds is taken, which tells us how many times higher or lower the mean frequency of the cells A=1 is than the mean frequency of the cells A=2. However, one-variable effects are seldom of interest because they tell nothing about the relationships between variables. More particularly, the values of the one-variable effects will change when additional variables are introduced. The amount of change reflects the importance of these additional variables for the model.

The *two-variable effect* refers to the average relationship between A and B within categories of the remaining variables. The effect AB, within a table ABC, measures the strength of the partial association between A and B, keeping C constant. The computation proceeds analogously to the calculation of the one-variable effects. First, a geometric mean is computed to determine the average size of the cell frequencies AB = (i,j). Second, it is

determined how many times this average frequency deviates from what was indicated by the lower order effects. In general, in any table with three or more variables with any number of category, the partial effect AB equals the geometric mean of all corresponding conditional effects AB.

The *three-variable effect* is a symmetric measure of interaction: the same value is always obtained regardless of which two-variable relationship is taken into account.

Extension to more variables follows easily. If effect coding is used, each partial effect τ can be calculated as the geometric mean of the corresponding conditional effects and each higher order effect can be calculated in terms of the extent to which the conditional effects deviate from the partial effect.

Yet, Kaufman and Schervish (1986, pp.718-722) expressed some concern about the use of odds ratio contrasts in the interpretation of results and about the potential danger of misinterpretation in general. In addition, readers who are not yet familiar with the concepts and the logic of odds and odds ratios may usually not understand fully the explanation of the findings. Moreover, Kaufman and Schervish extended the utility of log-linear analysis by presenting a new way of interpreting its parameters. The method consists in creating adjusted cross-tabulations. More specifically, this Deming-Stephan adjustment maintains the relationship between the variables, as indexed by the odds ratios. Yet, it adjusts the absolute cell sizes to match specified marginal counts, i.e. the number of cases in each category of the dependent variable and the number of cases in each category of the independent variable. Subsequently, the presentation of net distributions and percentages make log-linear results and effects easier for the researcher to understand and present and easier for readers to comprehend. Whereas log-linear parameters and odds ratios always express relative differences in effects, this new method provides a way to estimate both these relative differences as well as the actual absolute effects embedded in the log-linear parameters (Kaufman & Schervish, 1986, pp.717-718).

4.2.2.2. Causality

4.2.2.2.1 Categorical Data

So far, no assumptions have been made about the level of measurement of the variables in log-linear models. In general, variables are treated as nominal level variables. Many methodologists have suggested the use of methods which treat ordinal level data as if they are measured at interval level. Although they call it an ordinary level approach, they make assumptions which make sense only at the interval level of measurement.

4.2.2.2.2. Effect Models And Causality

Causality implies a straightforward classification of causes and consequences. However, as for log-linear modelling, and more specifically effect modelling, one should exercise caution in talking about causality. This can be illustrated by a statement from Hagenaars (1990, p.70): "This implicit preference to give the variables *a causal order* and to interpret their relationships *in a causal sense* can be expressed explicitly in effect models. […] The effect parameters of a logit model have *a causal connotation*." So, causality becomes a suspicious term in log-linear effect models.

Indeed, the question arises if relationships captured in models can be specified and interpreted in terms of causality. Although the issue of 'causality' can easily become mainly philosophical, one often deals with it by explaining the three criteria as described by J.S. Mill (in Meerling, 1995): A will cause B if (1) there exists covariance between A and B, (2) A proceeds in time, and (3) there is no other plausible explanation for the relationship between A and B.
In order to see if the first criterion holds, a model, once constructed, can be screened by matching it to the observations in de dataset (fitting a model). Yet, a good fit can never 'prove' the presence of a causal link in the relationships that are captured in the model. One can only state that the model is *plausible and not in conflict with the observations*. As such, even regression coefficients in se can not be interpreted in terms of causality. In regression analysis, an equation shows how one variable in best terms can be predicted out of other variables. Each variable can be deduced to some extent from other variables if they are correlated. In sum, correlation does not imply causality. Therefore, it will be the attributed and theoretical meaning and the inherently postulated hypotheses that will show which model can be 'treated' as a causal model and which can not. If model screening results into a good data fit and if there is a reasonable explanation for the empirically detected correlation between the manifest variables, one can interpret for example regression coefficients in terms of *(relative) causality*. However, Boomsma (1998) argues that in such a case, one should not speak about 'causality', whether relative or not, because in practice, it easily slips into causality stricto sensu. Therefore, another term is presented in order to avoid misinterpretation of empirical modelling in general. One should rather use the term *directional correlation,* or 'gerichte samenhang'.

4.3. Conclusions

In general, the design, the execution and the analysis of multi-national surveys present serious challenges. Indeed, the IALS has been the first large-scale comparative assessment of adult literacy skills and is characterised by multi-disciplinary measurement techniques and international effort on a grand scale.
Therefore, some extenuating circumstances have to be taken into account. First, contrary to the IALS, comparable previous international research often uses school populations. Consequently, sample procedures then are more transparent and response rates can be mapped more systematically. In addition, a school-setting generally evokes obligatory research participation. Second, the international structural organisation of the IALS is inherently confronted with some severe shortcomings. Financial resources needed to enter the IALS are gathered nationally. Thus, next to the scientific coordination centre, a policy principal is also involved. Throughout, more diversity is being created because the very participation in the IALS which is characterised by the directives and procedures of the ETS and Statistics Canada, is based on a certain voluntarism.

Despite the methodological imperfections, the instruments developed for measuring adult literacy constitute an important advance. Their results are considered a valuable contribution to the field. Moreover, in establishing an independent review of the methodology, the international coordination and survey team prove they are setting higher standards than have been employed in some previous international surveys.

Finally, a statistical technique is chosen. The nature of the IALS data (cross-sectional with mainly categorical data) directs the choice towards contingency table analysis. Its main goal is the comparative analyses of Flanders and the Netherlands. Log-linear modelling is found to be a useful frame for uncovering complex relationships between variables. The BMDP

software programme follows a conditional testing procedure of hierarchical, unsaturated models. Logit models are used in order to construct modified regression or path models. Yet, interpreting these models and the estimated parameters remains a cumbersome aspect. Moreover, although 'effect' models connote a certain causality with regard to the kind of relationship, causality stricto sensu can not be applied here. Researchers into log-linear models are with good reason very cautious about dealing with causality. In sum, it is better to interpret relations in log-linear models in terms of *directional correlation*. Chapter Five reports on the comparative log-linear modelling procedures in the analyses between Flanders and the Netherlands.